B

Challenge Problems in Bioinformatics and Computational Biology from Other Reports

B.1 GRAND CHALLENGES IN COMPUTATIONAL BIOLOGY (David Searls)[1]

1. Protein structure prediction
2. Homology searches
3. Multiple alignment and phylogeny construction
4. Genomic sequence analysis and gene-finding

B.2 OPPORTUNITIES IN MOLECULAR BIOMEDICINE IN THE ERA OF TERAFLOP COMPUTING (Klaus Schulten et al.)[2]

1. Study protein-protein and protein-nucleic acid recognition and assembly
2. Investigate integral functional units (dynamic form and function of large macromolecular and supramolecular complexes)
3. Bridge the gap between computationally feasible and functionally relevant time scales
4. Improve multiresolution structure prediction
5. Combine classical molecular dynamics simulations with quantum chemical forces
6. Sample larger sets of dynamical events and chemical species
7. Realize interactive modeling
8. Foster the development of biomolecular modeling and bioinformatics
9. Train computational biologists in teraflop technologies, numerical algorithms, and physical concepts
10. Bring experimental and computational groups in molecular biomedicine closer together.

[1] D. Searls, "Grand Challenges in Computational Biology," *Computational Methods in Molecular Biology*, S. Salzberg, D. Searls, and Simon Kasif, eds., Elsevier Science, 1998.

[2] K. Schulten, G. Budescu, F. Molnar, *Opportunities in Molecular Biomedicine in the Era of Teraflop Computing*, NIH Resource for Macromolecular Modeling and Bioinformatics, March 3-4, 1999, Rockville, MD; see http://whitepapers.zdnet.co.uk/0,39025945,60014729p-39000617q,00.htm.

B.3 WORKSHOP ON MODELING OF BIOLOGICAL SYSTEMS (Peter Kollman and Simon Levin)[3]

Challenging Issues That Span All Areas of Modeling Systems

A. Integrating data and developing models of complex systems across multiple spatial and temporal scales

- Scale relations and coupling
- Temporal complexity and coding
- Parameter estimation and treatment of uncertainty
- Statistical analysis and data mining
- Simulation modeling and prediction

B. Structure-function relationships

- Large and small nucleic acids
- Proteins
- Membrane systems
- General macromolecular assemblies
- CeIlular, tissue, organismal systems
- Ecological and evolutionary systems

C. Image analysis and visualization

- Image interpretation and data fusion
- Inverse problems
- Two-, three- and higher-dimensional visualization and virtual reality

D. Basic mathematical issues

- Formalisms for spatial and temporal encoding
- Complex geometry
- Relationships between network architecture and dynamics
- Combinatorial complexity
- Theory for systems that combine stochastic and nonlinear effects often in partially distributed systems

E. Data management

- Data modeling and data structure design
- Query algorithms, especially across heterogeneous data types
- Data server communication, especially peer-to-peer replication
- Distributed memory management and process management

B.4 WORKSHOP ON NEXT-GENERATION BIOLOGY: THE ROLE OF NEXT-GENERATION COMPUTING (Shankar Subramaniam and John Wooley)[4]

Exemplar Challenges for Bioinformatics and Computational Biology

1. Full genome-genome comparisons
2. Rapid assessment of polymorphic genetic variations

[3] "Modeling of Biological Systems," P. Kollman and S. Levin (chairs), a workshop at the National Science Foundation, March 14 and 15, 1996, available at http://www.resnet.wm.edu/~jxshix/math490/Modeling%20of%20Biological%20Systems.htm.

[4] S. Subramaniam and J. Wooley, DOE-NSF-NIH 1998 Workshop on Next-Generation Biology: The Role of Next Generation Computing, available at http://cbcg.lbl.gov/ssi-csb/nextGenBioWS.html.

3. Complete construction of orthologous and paralogous groups of genes
4. Structure determination of large macromolecular assemblies/complexes
5. Dynamical simulation of realistic oligomeric systems
6. Rapid structural/topological clustering of proteins
7. Prediction of unknown molecular structures; protein folding
8. Computer simulation of membrane structure and dynamic function
9. Simulation of genetic networks and the sensitivity of these pathways to component stoichiometry and kinetics
10. Integration of observations across scales of vastly different dimensions and organization to yield realistic environmental models for basic biology and societal needs

B.5 TECHNOLOGIES FOR BIOLOGICAL COMPUTER-AIDED DESIGN (Masaru Tomita)[5]

1. Enzyme engineering: to refine enzymes and to analyze kinetic parameters in vitro
2. Metabolic engineering: to analyze flux rates in vivo
3. Analytical chemistry: to determine and analyze the quantity of metabolites efficiently
4. Genetic engineering: to cut and paste genes on demand, for modifying metabolic pathways
5. Simulation science: to efficiently and accurately simulate a large number of reactions
6. Knowledge engineering: to construct, edit and maintain large metabolic knowledge bases
7. Mathematical engineering: to estimate and tune unknown parameters

B.6 TOP BIOINFORMATICS CHALLENGES (Chris Burge et al.)[6]

1. Precise, predictive model of transcription initiation and termination: ability to predict where and when transcription will occur in a genome
2. Precise, predictive model of RNA splicing/alternative splicing: ability to predict the splicing pattern of any primary transcript
3. Precise, quantitative models of signal transduction pathways:ability to predict cellular response to external stimuli
4. Determining effective protein-DNA, protein-RNA and protein-protein recognition codes
5. Accurate ab initio structure prediction
6. Rational design of small molecule inhibitors of proteins
7. Mechanistic understanding of protein evolution: understanding exactly how new protein functions evolve
8. Mechanistic understanding of speciation: molecular details of how speciation occurs
9. Continued development of effective gene ontologies-systematic ways to describe the functions of any gene or protein
10. (Infrastructure and education challenge)
11. Education: development of appropriate bioinformatics curricula for secondary, undergraduate, and graduate education

B.7 EMERGING FIELDS IN BIOINFORMATICS (Patricia Babbitt)[7]

1. Data storage and retrieval, database structures, annotation
2. Analysis of genomic/proteomic/other high-throughput information

[5] M. Tomita, "Towards Computer Aided Design (CAD) of Useful Microorganisms," *Bioinformatics* 17(12):1091-1092, 2001.

[6] C. Burge, "Bioinformaticists Will Be Busy Bees," *Genome Technology*, No. 17, January, 2002. Available (by free subscription) at http://www.genome-technology.com/articles/view-article.asp?Article=20021023161457.

[7] P. Babbitt et al., "A Very Very Very Short Introduction to Protein Bioinformatics," August 22-23, 2002, University of California, San Francisco, available at http://baygenomics.ucsf.edu/education/workshop1/lectures/w1.print2.pdf.

3. Evolutionary model building and phylogenic analysis
4. Architecture and content of genomes
5. Complex systems analysis/genetic circuits
6. Information content in DNA, RNA, protein sequences and structure
7. Metabolic computing
8. Data mining using machine learning tools, neural nets, artificial intelligence
9. Nucleic acid and protein sequence analyses

B.8 TEN GRAND CHALLENGES (Sylvia Spengler)[8]

1. The origin, structure, and fate of the universe
2. The fundamental structure of matter
3. Earth's physical systems
4. The diversity of life on Earth
5. The tree of life
6. The language of life
7. The web of life
8. Human ecology
9. The brain and artificial thinking machines
10. Integrating Earth and human systems
11. A knowledge server for planetary management

Research Across Domains: Data

- Information management—human evolution continued
- Exponential increase in data and information across domains
- Access to information across domains—as or more important than the information itself
- Integration of data across knowledge domains
- Apply analytical tools across knowledge domains
- Modeling of complex systems
- Simulation of phenomena—descriptive science becomes predictive science

Research Across Domains: People

- Share data across disciplines
- Build and use analytical and modeling tools across disciplines
- Work in collaborative, cross-domain groups

Research Across Domains: Time

- Real-time data access, integration, and analysis
- Real-time modeling and effects prediction
- Real-time dissemination of research results
- Real-time testing by research community
- Real-time policy discussions
- Real-time policy decisions

[8] S. Spengler, Lawrence Berkeley National Laboratory, personal communication to John Wooley, January 3, 2005.

B.9 GRAND CHALLENGES IN BIOMEDICAL COMPUTING (John A. Board, Jr.)[9]

Biomedical Applications from Coupling Imaging and Modeling

- Real-time noninvasive three-dimensional imaging of many body systems
- Real-time generation of three-dimensional patient-specific models
- Multiple-technology (multimodal) imaging and modeling
- Whole-organ modeling
- Multiple-organ system modeling
- Patient-specific modeling of organ anomalies
- Model support for (partial) restoration of hearing, coarse vision, and locomotion (via both paralyzed and artificial limbs)

All of these applications make use of:

- Three-dimensional models
- Increasingly refined grids and increasing levels of tissue discrimination
- Anatomically realistic models
- Special-purpose hardware for visualization
- Distributed computing techniques.

B.10 ACCELERATING MATHEMATICAL-BIOLOGICAL LINKAGES: REPORT OF A JOINT NSF-NIH WORKSHOP (Margaret Palmer et al.)[10]

List of Top Ten Problems at the Mathematical Biology Interface

1. Model multilevel systems: from the cells in people, to human communities in physical, chemical, and biotic ecologies.
2. Model networks of complex metabolic pathways, cell signaling, and species interactions.
3. Integrate probabilistic theories: understand uncertainty and risk.
4. Understand computation: gaining insight and proving theorems from numerical computation and agent-based models.
5. Provide tools for data mining and inference.
6. Address linguistic and graph theoretical approaches.
7. Model brain function.
8. Build computational tools for problems with multiple temporal and spatial scales.
9. Provide ecological forecasts.
10. Understand effects of erroneous data on biological understanding.

B.11 GRAND CHALLENGES OF MULTIMODAL BIOMEDICAL SYSTEMS (J. Chen et al.)[11]

Science Challenges

1. Allow early detection of where and when an infectious disease outbreak occurs, whether it is naturally occurring or man-made, in real time.

[9] J.A. Board, Jr., "Grand Challenges in Biomedical Computing, *High-Performance Computing in Biomedical Research*, T.C. Pilkington, B. Loftis, J.F. Thompson, S.L.Y. Woo, T.C. Palmer, and T.F. Budinger, eds., CRC Press, Boca Raton, FL, 1993.

[10] M. Palmer et al., "Accelerating Mathematical-Biological Linkages: Report of a Joint NSF-NIH Workshop," February 2003, available at www.maa.org/mtc/NIH-feb03-report.pdf.

[11] J. Chen et al., "Grand Challenges of Multimodal Bio-Medical Systems," *IEEE Circuits and Systems Magazine*, pp. 46-52, 2nd Quarter 2005, available at http://gsp.tamu.edu/Publications/PDFpapers/pap_CASmag_MBM.pdf.

2. Develop multidimensional drug profiling databases to facilitate drug discovery and to identify biomarkers for diagnosis and monitoring the progress of individual disease treatments.
3. Connect activities and events derived from cellular processes to high-level cognitions.
4. Support personalized medical care and clinical decision for patients

Technology Challenges and Enabling Technologies

1. Formalization of biological knowledge into predictive models for systems biology and system-based analysis
2. Interdisciplinary training
3. Development of open source, multiscale modality informatics toolkits

B.12 THE DEPARTMENT OF ENERGY'S GENOMES TO LIFE PROGRAM[12]

21st Century Biology Requiring "Biocomp" Tools

1. Population models, symbiosis, and stability
2. Discrete growth models
3. Reaction kinetics
4. Biological oscillators and switches
5. Coupled oscillators
6. Reaction-diffusion, chemotaxis, and nonlocality
7. Oscillator-generated wave phenomena and patterns
8. Spatial pattern formation with population interactions
9. Mechanical models for generating pattern and form in development
10. Evolution and morphogenesis

A Mathematica for Molecular, Cellular, and Systems Biology

1. Core data models and structures [database management]
2. Optimized functions [core libraries]
3. Scripting environment [e.g., Python, PERL, ruby, etc.]
4. Database accessors and built-in schemas
5. Simulation interfaces
6. Parallel and accelerated kernels
7. Visualization interfaces (for information visualization and scientific visualization)
8. Collaborative workflow and group use interfaces

Hierarchical Biological Modeling Environment

1. Genetic sequences
2. Molecular machines
3. Molecular complexes and modules
4. Networks + pathways [metabolic, signaling, regulation]
5. Structural components [ultrastructures]
6. Cell structure and morphology
7. Extracellular environment
8. Populations and consortia

[12] R. Stevens, "GTL Software Infrastructure: A Computer Science Perspective," undated presentation, Argonne National Laboratory, available at www.doegenomics.org/compbio/mtg_1_22_02/RickStevens.ppt.

Modeling and Simulation Challenges for 21st Century Biology

1. Modeling activity of single genes
2. Probabilistic models of prokaryotic genes and regulation
3. Logical models of regulatory control in eukaryotic systems
4. Gene regulation networks and genetic network inference in computational models and applications to large-scale gene expression data
5. Atomistic-level simulation of biomolecules
6. Diffusion phenomena in cytoplasm and extracellular environment
7. Kinetic models of excitable membranes and synaptic interactions
8. Stochastic simulation of cell signaling pathways
9. Complex dynamics of cell cycle regulation
10. Model simplification

B.13 HIGH-PERFORMANCE COMPUTING, COMMUNICATION, AND INFORMATION TECHNOLOGY GRAND CHALLENGES (LATE 1980s, EARLY 1990s)[13]

Computing Applications to Map and Sequence Human Genome

1. Understanding protein folding
2. Predicting structure of native protein
3. Exhaustive discovery and analysis of cancer genes
4. Molecular recognition and dynamics
5. Drug discovery

[13] Committee on Physical, Mathematical, and Engineering Sciences of the Federal Coordinating Council for Science, Engineering, and Technology, U.S. Office of Science and Technology Policy, FY1992 Blue Book: *Grand Challenges: High Performance Computing and Communications—The FY 1992 U.S. Research and Development Program.*